ELAN : Una Red Distribuida y poliglota de recursos linguisticos textuales

Samuel Cruz-Lara

ELAN : UNA RED DISTRIBUIDA Y POLÍGLOTA DE RECURSOS LINGÜÍSTICOS TEXTUALES

Samuel CRUZ-LARA

Laboratoire Lorrain de Recherche en Informatique et ses Applications

LORIA (UMR 7503) CNRS – INRIA – Universités de Nancy

http://www.loria.fr

Equipe « Langue et Dialogue »

Nancy, FRANCE

Samuel.Cruz-Lara@loria.fr

Resumen. En este artículo presentamos el proyecto MLIS-ELAN cuyo objetivo principal es la creación de una red distribuida y políglota de bases de datos de recursos lingüísticos textuales. Primero, definimos el término de "recurso lingüístico textual" y mostramos la importancia de la normalización en el marco de la representación y de la utilización de este tipo de recursos. Después, presentamos los objetivos del proyecto y mostramos la "arquitectura de software" que ha permitido alcanzar dichos objetivos. Particularidad importante del modelo propuesto: las bases de datos que contienen los recursos lingüísticos textuales son heterogéneas y se encuentran físicamente en sitios diferentes. Los usuarios, sin embargo, tienen la impresión de utilizar una base de datos única además de disponer de herramientas que les permiten consultar, comparar y almacenar los recursos lingüísticos elegidos, independientemente de su formato y de su localización física.

1 LOS RECURSOS LINGÜÍSTICOS TEXTUALES

En este documento, el término "recurso lingüístico textual" incluye todo dato textual que provenga de una fuente escrita (novelas, periódicos, revistas, ...), o bien, de una fuente oral (transcripción de diálogos, por ejemplo). Debe entonces entenderse por "recurso lingüístico textual", todo tipo de objeto, documento o soporte, que se use en el marco del estudio o del tratamiento del lenguaje oral o escrito.

Cabe aclarar del mismo modo, que el término "recurso lingüístico textual" hace referencia a datos de tipo "multimodal". Es decir, que se incluye, obviamente, el componente textual clásico (corpus, glosarios, ...), pero también se incluyen datos de tipo temporal, espacial (el gesto), referencial (objetos y acciones) sin olvidar la dimensión multimedia (sonido e imagen).

2 CODIFICACIÓN Y NORMALIZACIÓN

La utilización de recursos lingüísticos textuales que provienen de informaciones heterogéneas, necesita la elaboración de métodos y de herramientas de creación, de manipulación y de gestión que pueden llegar hasta la construcción de un verdadero entorno de edición y de tratatimiento de corpus "multimodales" (p.e. palabra y gesto). Dichos métodos, inspirados directamente en la ingeniería del documento electrónico, se orientan hacia una gestión de recursos lingüísticos textuales fuertemente normalizados y codificados según los estándares internacionales tales como SGML[1]/TEI[2]/XML[3].

La normalización es, en efecto, muy importante en el marco de la gestión de recursos lingüísticos textuales, puesto que ésta les permite, a la vez, ser independientes de toda aplicación de software y permanecer, sin embargo, fácilmente intercambiables y comparables con otros recursos [1], [2], [3].

La experiencia demuestra que una de las dificultades más importantes, en lo que a la manipulación de textos se refiere, es la selección del tipo de codificación utilizado. De esta selección depende, en gran parte, el éxito de un estudio, de una investigación lingüística, del almacenamiento o del intercambio de todo recurso lingüístico textual.

Si se excluye la "World Wide Web", el soporte puede ser tan variado como un disquete, una cinta magnética o un CD-ROM. El intercambio, obviamente, debe entonces efectuarse en las mejores condiciones posibles para no perder informaciones ni deteriorar la estructura de los datos. La selección del tipo de codificación se orientará entonces hacia un formato fácilmente transmisible de un soporte a otro, de una arquitectura a otra.

Sea cual sea la forma en la que el texto se presenta, éste no sera utilizable si no se dispone de herramientas para su manipulación. Un cierto nivel de normalización permitirá poner a disposición de cualquier usuario un conjunto de herramientas estandarizadas y evitará la duplicación de herramientas que efectúen la misma función sobre formatos diferentes, algunas veces exóticos.

Desde hace ya varios años, el mercado dispone de bibliotecas, de herramientas y más aún, de entornos de programación completos que soportan, completa o parcialmente, la norma SGML.

---

[1] Standard Generalized Markup Language (SGML).
International Standards Organization, ISO 8879: Information processing---Text and office systems---Standard Generalized Markup Language (SGML), ([Geneva]: ISO, 1986)

[2] Guidelines for the encoding and interchange of machine-readable texts edited by C.M.Sperberg-McQueen and Lou Burnard (Chicago and Oxford, ALLC-ACH-ACL Text Encoding Initiative, 1994)

[3] XML : eXtended Markup Language W3C REC-xml-19980210, XPointer : XML Pointer Language W3C WD-xptr-19980303, XLink : XML Linking Language WD-xlink-19980303

La norma SGML es de hecho muy utilizada en el entorno de la edición y de la gestión documentaria empresarial, pero su uso se revela pesado y muy complejo. Esto ha limitado fuertemente su utilización, particularmente, en los medios académicos.

Por ello, la TEI "Text Encoding Initiative", que es de hecho una aplicación "aligerada" de SGML, proporciona un mecanismo riguroso y eficaz de codificación de recursos lingüísticos textuales.

Las formas textuales que pueden ser codificadas gracias a la TEI son muy numerosas y polivalentes, por ejemplo, pueden codificarse novelas, teatro, artículos técnicos o científicos, pero también diccionarios y diálogos "multimodales".

Una característica fundamental de la TEI es que, al contrario de HTML, que es también una aplicación de SGML, la TEI normaliza la etapa de representación de los datos.

Otra ventaja de la TEI, además de la codificación de texto propiamente dicha, es que dispone de un "encabezado". El encabezado es, sin duda, el punto fuerte de la TEI, ya que indica de manera muy precisa la versión electrónica del texto, el tipo de codificación utilizado, el origen del texto fuente (bibliografía) y las modificaciones relativas a la codificación desde la creación del formato electrónico.

Para ser más precisos, consideremos el caso de documentos textuales que provienen de una fuente escrita (novelas, periódicos, ...) o de una fuente oral (transcripciones de diálogos). Los usuarios de estos documentos deben poder buscar los recursos que necesitan eficientemente, y una vez encontrados, poder manipularlos fácilmente. Para que este objetivo sea alcanzado, es obvio que es necesaria una representación normalizada de esos recursos.

Las recomendaciones de la TEI tienen por objeto facilitar la utilización de SGML a través del uso de marcos normativos para la codificación de documentos. La polivalencia de los recursos lingüísticos utilizados por el proyecto "SILFIDE" [4] refleja bien todas las posibilidades de la TEI :

- obras literarias de tipo novela,
- comics,
- textos oficiales provenientes de la Union Europea,
- diccionarios,
- artículos de periódicos y
- transcripciones de diálogos.

Sin embargo, las posibilidades que la TEI ofrece son tan amplias, que sin una política rigurosa de codificación, se corre el riesgo de ver aparecer una gran disparidad en lo que a la representación de los contenidos se refiere. Luego entonces, es importante que se proponga una maqueta común a todos los textos similares en contenido y forma, a fin de garantizar a los usuarios un contenido mínimo y la obtención de resultados.

XML, que fue apenas creado en 1998, es también una aplicación de SGML cuyo objetivo fundamental es el de abrir las puertas de Internet a los documentos semi-estructurados, además de que se superan diversas deficiencias de HTML. XML provee un método riguroso, eficaz y muy simple para la descripción y el intercambio de información estructurada en la Web. XML establece de hecho un justo equilibrio entre la complejidad de SGML y la pobreza semántica de HTML. XML es un subconjunto de SGML, HTML es una aplicación de SGML.

El consorcio TEI, recientemente creado, modifica actualmente las directivas de la TEI para adaptarlas a la recomendación XML[4].

Aún cuando comparar XML y HTML no es el objetivo de este documento, podemos mencionar, que un documento XML preserva la semántica y la estructura de los datos.

No obstante, podemos mencionar igualmente que, a diferencia de HTML, XML no da ningún detalle referente a la presentación de un documento (*tag*
, por ejemplo). En efecto, XML describe únicamente la estructura lógica de un documento, sin ocuparse, *a priori*, de la manera en la que éste se imprimirá o simplemente se mostrará a los usuarios.

XSL[5] es una aplicación de XML que permite la definición de "hojas de estilo", a través de las cuales cualquier documento XML es formateado, ya sea con el objetivo de imprimirlo, o mostrarlo. Así, por ejemplo, a partir de un documento XML único, se pueden obtener, a través del uso de varias "hojas de estilo" diversos tipos de visualización, de edición o de impresión. Además, XSL permite la transformación de un documento XML en otro documento XML o inclusive en un documento HTML que puede ser visualizado en cualquier navegador web. XSL es de hecho mucho más que un "simple" lenguaje de "estilo".

XSL está constituido por dos lenguajes :

1. XSLT[6] que es un poderoso lenguaje de transformación.
2. Un vocabulario XML que permite la especificación de la semántica de presentación.

3 EL PROYECTO MLIS-ELAN.

El objetivo del proyecto ELAN "European Language Activity Network", proyecto de MLIS "Multi Lingual Information Society" [5], [6], es definir una arquitectura que permita la utilización de grandes fondos europeos de recursos lingüísticos textuales.

---

[4] Se podrá consultar http://www.tei-c.org/ para obtener más detalles.
[5] XSL : eXtensible Stylesheet Language W3C Working Draft 12 January 2000
[6] XSLT : XSL Transformations W3C REC-xslt-19991116

El proyecto MLIS-ELAN está estructurado alrededor de dos grandes temas :

1. La definición de un lenguaje de consulta común a todos los servidores de le red, CQL “Common Query Language”.
2. La construcción de una arquitectura distribuida que permita a los usuarios acceder a grandes bases de datos repartidas en diversos lugares distantes entre sí[7].

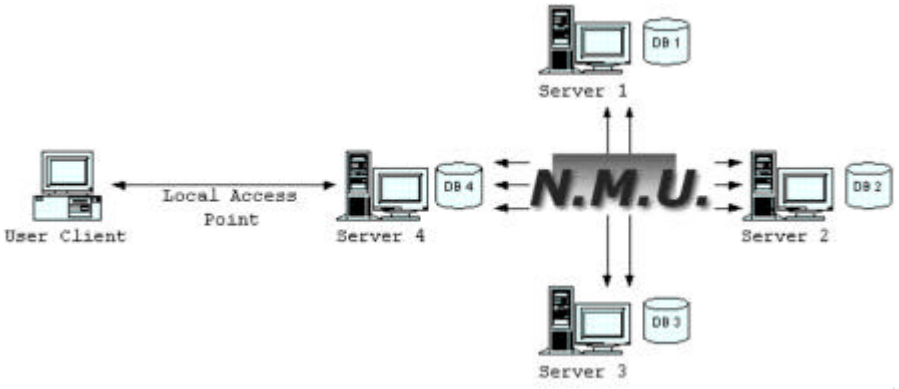

Figura 1. Arquitectura general del proyecto MLIS-ELAN.

En la figura 1 hemos representado una serie de servidores. Cada servidor posee su propia base de recursos lingüísticos. Cada una de esas bases utiliza un sistema de codificación normalizado y basado en la TEI. Sin embargo, debe aclararse que dichas bases de datos son heterogéneas, es decir, que por un lado puede haber una base de datos relacional y por otro lado puede haber una base de datos orientada a objeto.

Un usuario se conecta a la red a través de un servidor “local” en el cual está debidamente registrado e identificado. Ya que cada servidor dispone de su propia base de recursos lingüísticos, es importante, para una utilización fácil, que los usuarios puedan acceder a todo el conjunto de recursos lingüísticos de la red, conservando la impresión de que todos esos recursos se encuentran físicamente en el servidor “local”. Las arquitecturas de SILFIDE y de MLIS-ELAN funcionan de esta manera [7].

Para que una arquitectura de este tipo pueda proveer a los usuarios los recursos y las herramientas que necesitan, hay que describir, a nivel de la arquitectura de software, el protocolo que será usado para el intercambio de datos (puede tratarse de recursos lingüísticos o de informaciones relativas, por ejemplo, al tráfico de la red) entre los diferentes servidores. Las comunicaciones entre los diferentes servidores de

[7] http://www.loria.fr/projets/MLIS/ELAN

la red son efectuadas por una aplicación de software llamada “Network Management Unit” (NMU). La NMU controla igualmente el acceso a todos los recursos lingüísticos y permite a cada usuario el almacenamiento de los recursos que le interesan.

La ventaja fundamental de XML es que se le puede usar igualmente para describir el protocolo utilizado por la NMU.

XML nos permite entonces normalizar, no solamente la representación de los recursos lingüísticos disponibles, sino también los protocolos usados para el intercambio de información entre los diferentes servidores de la red.

En este contexto, XML debe ser considerado como un lenguaje de descripción de alto nivel utilizado en particular por la NMU. En lo que se refiere a lo que podríamos llamar el “nivel bajo”, nos apoyamos sobre una plataforma de tipo CORBA[8].

Además, XML permite normalizar igualmente la manipulación de recursos lingüísticos. Hemos definido, por un lado, un lenguaje de interrogación (“Query Language”) que permite a los usuarios describir las características de los recursos lingüísticos que desean utilizar y por otro lado, a nivel “sistema”, hemos descrito la forma en la que los recursos seleccionados serán presentados a los usuarios (“Result Set”).

Una sesión de trabajo con MLIS-ELAN se desarrolla de la siguiente forma :

1. Conexión del usuario a su servidor local. Para conectarse a la red MLIS-ELAN un navegador web[9] es suficiente. Obviamente, para poder conectarse, el usuario debe ser miembro de la red. La pertenencia es verificada mediante un mecanismo clásico de tipo “palabra clave”. A partir del momento en que el usuario ha sido debidamente identificado, la NMU se encarga de establecer la comunicación con los demás servidores de la red. Cabe señalar que todos los elementos de la interfaz de usuario existen en diversos idiomas. Por lo pronto, francés, inglés, alemán y holandés. A corto plazo, todos los idiomas “oficiales” de la Union Europea podrán ser utilizados.

---

[8] CORBA : Common Object Request Broker Architecture. OMG (Object Management Group) http://www.omg.org

[9] Es necesario utilizar un “browser” internet compatible Java 2 (jdk1.2 mínimo), o bien, compatible jdk 1.1.6 mínimo con Swing 1.1.

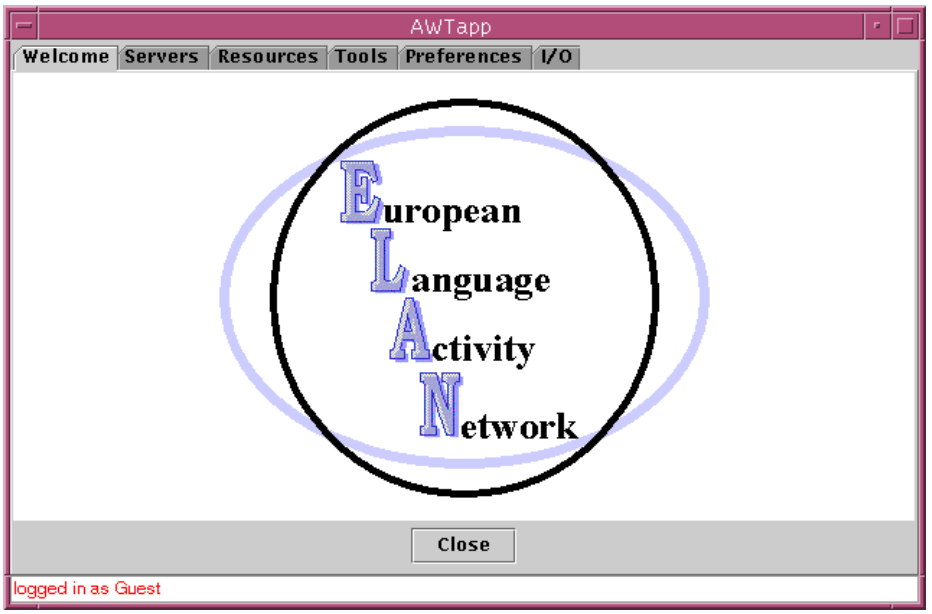

Figura 2. La interfaz de usuario de MLIS-ELAN.

2. La segunda etapa consiste en seleccionar la lista de servidores con los que se desea trabajar. Esta selección se hace simplemente a través de una lista. El primer prototipo del proyecto MLIS-ELAN incluía servidores de recursos lingüísticos de la Universidad de Birmingham (Reino Unido), de la Universidad de Pisa (Italia), del Instituto de Lexicología de Leiden (Holanda) y del Instituto Nacional de la Lengua Francesa de Nancy (Francia).

3. Selección de un subconjunto de recursos. Esta selección se hace bajo la forma de un proceso iterativo de peticiones que se envían a los servidores que se han escogido. El usuario crea de esta manera un "corpus virtual", es decir, un subconjunto de los corpus que se desea explorar. Una vez que dicho "corpus virtual" ha sido creado, es posible, mediante el uso del lenguaje CQL, efectuar búsquedas a partir del nombre de un autor, de un título de obra, ... Obviamente, otro tipo de búsquedas más sofisticado es posible.

4. Acceso a los documentos. El usuario, a través de peticiones, selecciona los documentos que le interesan en el "corpus virtual". Estas peticiones son igualmente formuladas a través del lenguaje CQL. Gracias a CQL es posible, por ejemplo, buscar en el conjunto de recursos seleccionados, todas las frases que contienen todas las ocurrencias de una u otra palabra. De la misma manera, CQL permite la comparación entre varios recursos (concordancias), aún y cuando éstos no utilizen la misma lengua.

Luego entonces, en el proyecto MLIS-ELAN dos tipos de peticiones son posibles :

- Peticiones de selección de recursos (“Result Set Query”). Este tipo de peticiones corresponde al punto número 3.
- Peticiones de acceso al contenido (“Content Access Query”). Este tipo de peticiones corresponde al punto número 4.

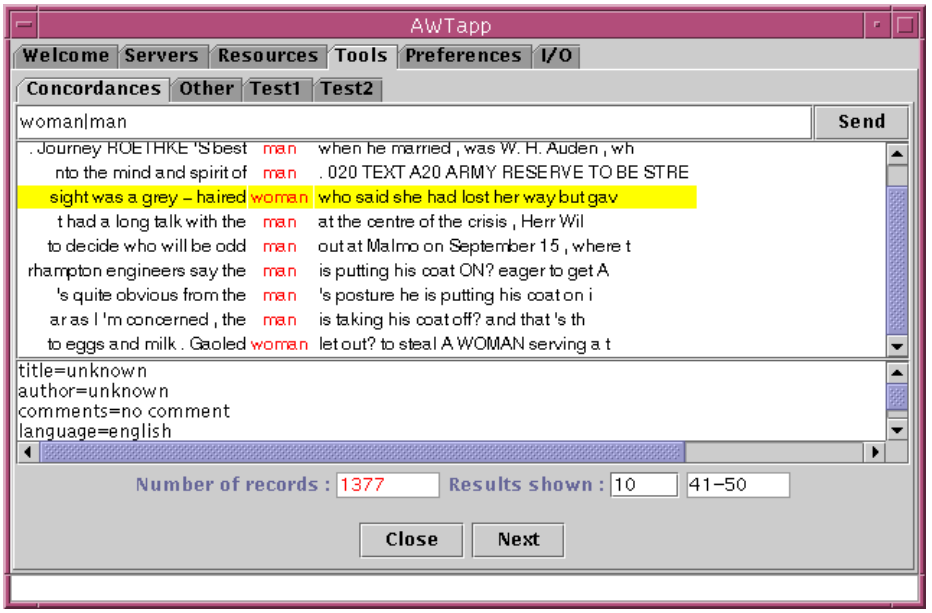

Figura 3. Resultado de una operación CQL.

5. Finalmente, es obviamente posible almacenar el “espacio de trabajo” del usuario. Éste incluye los servidores y los recursos seleccionados, así como los resultados de las diversas operaciones efectuadas, por ejemplo, las concordancias. El espacio de trabajo del usuario es controlado por la NMU.

4 CONCLUSIÓN.

¿Cuál es la originalidad del proyecto MLIS-ELAN? ¿Es el hecho de codificar los recursos lingüísticos textuales en XML/TEI? ¿Es el hecho de proponer una herramienta de búsqueda de información de tipo cliente/servidor?

Pues bien, ¡creemos que ninguno de esos aspectos es realmente novedoso! , aún y cuando MLIS-ELAN haya sido uno de los primeros proyectos a utilizar la forma XML de la TEI. Analicemos entonces cuáles son las características que hacen que MLIS-ELAN sea un proyecto realmente original.

Antes que nada, y tal y como se ha mencionado en el artículo, se debe tomar en cuenta que las bases de recursos lingüísticos de la red son heterogéneas. Lo cual quiere decir que cada base de recursos utiliza el modelo que le conviene. Puede haber

bases de datos relacionales, jerárquicas, orientadas a objeto o de cualquier otro tipo. La originalidad de la arquitectura diseñada para MLIS-ELAN reside en proponer, por encima del modelo que cada base de datos utiliza, un modelo de más alto nivel que es común a todas las bases de le red.

Las características más importantes de dicho modelo son las siguientes :

1. Un usuario de la red MLIS-ELAN se conecta directamente a su servidor local. Una vez que el usuario ha sido debidamente identificado, éste puede comenzar su búsqueda. La NMU controla la comunicación con los demás servidores de la red y el acceso a los recursos lingüísticos.
2. Todos les recursos lingüísticos que un usuario ha seleccionado, o bien, todos los recursos lingüísticos que resultan de una consulta de tipo CQL, son codificados en XML/TEI. El uso de la TEI en particular, permite la identificación de los diferentes elementos que constituyen un documento (p.e. párrafo, capítulo, sección, etc), pero la TEI permite igualmente la comparación entre dos o más recursos (concordancias).
3. El protocolo de alto nivel utilizado por la NMU está igualmente basado en XML. El protocolo de bajo nivel, está basado en CORBA.
4. Quada claro, igualmente, que otra originalidad del modelo propuesto es que la arquitectura de MLIS-ELAN pueder ser utilizada en otras áreas. Actualmente, en el marco de una "llamada de oferta" relativa a un proyecto de bio-informática, hemos propuesto que el modelo MLIS-ELAN sea utilizado.

Para más información relativa al proyecto MLIS-ELAN se puede consultar (http://solaris3.ids-mannheim.de/elan/index.html), o bien, para conocer los detalles más técnicos del proyecto (http://www.loria.fr/projets/MLIS/ELAN).

Finalmente señalemos que la "Comunidad ELAN" incluye en España[10] :

- Institut d'Estudis Catalans, Barcelona
- Fundació Bosch Gimpera, Universitat de Barcelona
- Real Academia de la Lengua Española, Instituto de Lexicografía

[10] La lista completa http://solaris3.ids-mannheim.de/elanbo1/elanbooklet/sites.html

5 BIBLIOGRAFÍA.

1. « Du document électronique à son usage : le rôle central de la normalisation ».
   J.-L. Benoît, Ch. Bernet, P. Bonhomme, L. Romary, N.Viscogliosi
   Revue « SOLARIS », Décembre 1999 / Janvier 2000. ISSN : 1265-4876.

2. « Les enjeux de la normalisation à l'heure du développement de l'information "dématérialisée" ».
   E. Giuliani
   Revue « SOLARIS », Décembre 1999 / Janvier 2000. ISSN : 1265-4876.

3. TEI-P3, Association for Computers and the Humanities (ACH), Association for Computational Linguistics.
   (ACL) and Association for Literary and Linguistic Computing (ALLC) 1994, Guidelines for Electronic Text Encoding and Interchange
   (TEI-P3), 2 vol., .Ed.C.M. Sperberg-McQueen and Lou Burnard,
   Chicago, Oxford : Text Encoding Initiative.

4. « The SILFIDE Network : An Interactive Service for Using, Studying, Distributing and Sharing Natural Language Resources »
   P. Bonhomme, S. Cruz-Lara and L. Romary
   SGML / XML ’97
   Washington D.C., USA. December 1997.

5. Deliverables 3.1-1 et 3.2-1 « User Client » and « Network Management Unit »
   C. de Saint-Rat, S. Cruz-Lara, P. Bonhomme and L. Romary
   MLIS (Multi-Lingual Information Society)
   ELAN Project (European Language Activity Network). December 1999.

6. Deliverable 3.3-1 « The ELAN Common Query Language (CQL) »
   Peter van der Kamp and Pieter Masereeuw
   MLIS (Multi-Lingual Information Society)
   ELAN Project (European Language Activity Network). December 1999.

7. « A General XML-based Distributed Software Architecture for Accessing and Sharing Resources »
   S. Cruz-Lara, P. Bonhomme, C. de Saint-Rat and L. Romary
   XML Finland ‘99
   Helsinki, Finlande. September 1999.